图书情报工作
第66卷 第10期 2022年5月

结合计量分析和内容分析的科学数据集使用特征研究*

■ 杨宁[1,2] 张志强[1,2]
[1]中国科学院成都文献情报中心 成都 610041
[2] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100190

摘 要：[目的/意义]从计量分析和内容分析两个视角对科学数据集的使用特征进行研究，定量化评估科学数据集对学科发展的影响，为科学数据管理服务及政策研究提供参考。[方法/过程]综合运用文本挖掘和文献计量方法对PubMed Central的全文文献进行分析，从时间分布、使用强度等7个方面全面考察科学数据集的使用情况，并在此基础上评估科学数据集对学科发展产生的实际影响。[结果/结论]研究结果表明，科学数据集对生物医学领域科研产生的影响力与日俱增，数据出版和高水平期刊促进了科学数据集的开放和共享，科学数据集的使用集中在论文的后半部分且正式引用较少，相应的标准规范还有待进一步加强。

关键词：计量分析 内容分析 科学数据集 使用特征
分类号：G203
DOI：10.13266/j.issn.0252-3116.2022.010.011

1 引言

科学数据集是科研活动过程中产生或经过再加工得到的，具有一定规范且可形成完整描述的数据资料或数据产品，主要类型包括实验数据、观测数据和统计数据等[1]。随着开放科学运动的兴起，科学数据集的共享和重用等使用行为变得日益普遍，其已逐渐成为贯穿科研全过程的重要研究对象和产出类型之一。对科学数据集使用特征和所产生的影响进行研究，一方面可以了解数据使用现状、掌握当前科研人员对数据的需求特征和利用情况；另一方面也可以具象并量化科学数据集对科研活动的实际贡献价值、合理规划科研资源配置、丰富科研评价指标。

当前，对科学数据集使用特征的研究一般采用计量分析或内容分析的方法。计量分析法是一种基于数学和统计学，以各种知识实体的外部特征和宏观特征为研究对象的定量分析方法[2]。从计量分析角度出发，一般采用数据集被引频次、使用下载量、被提及次数等指标对其使用特征和影响力进行研究评价。C. W. Belter等[3]以海洋学领域数据集为研究对象，利用被引次数研究数据集引用行为并对数据集的影响力进行评估。焦红等[4]运用文献计量方法从多维度对生物医学领域科学数据集的复用特征规律进行研究，并对高频复用数据集进行了详细分析。计量分析可以从宏观的角度对学科领域科学数据集的使用情况进行分析，进而度量数据集对整个学科发展产生的影响力；内容分析法则深入到学术论文全文本内容层面，通过人工判读或自然语言处理等方法研究文献内隐含的各种知识实体的使用行为特征[5]。从内容分析角度出发，一般通过数据集的使用方式、使用位置、使用强度等指标研究其使用特征和影响力。王雪等[6]以CNKI中10个学科的文献为研究对象，采用内容分析法从数据提及方式、使用位置、来源类型等角度分析比较了不同学科数据重用行为的特征。李龙飞等[7]从替代计量学视角出发，以地球系统科学数据共享平台的数据集为研究对象，利用内容分析方法对科学数据集使用方式进行研究并定量测度其价值。内容分析法的分析层面更加微观，可以从细粒度的文章结构层面研究数据集使用特征及影响力。从当前相关研究的现状来看，由于科学数据集使用特征的学科差异性较大、文献中科学数据集信息难以识别和抽取等问题，对科学数据集使用特征和产生影响的研究还多采用人工标注或围绕小范围数据开展，分析层面和分析指标也较为宽泛。

本研究将以生物医学领域大规模学术论文集作为

* 本文系国家社会科学基金重点项目“面向领域知识发现的学科信息学理论与应用研究”（项目编号：17ATQ008）研究成果之一。
作者简介：杨宁，副研究馆员，博士研究生；张志强，研究员，博士生导师，通信作者，E-mail：zhangzq@clas.ac.cn。
收稿日期：2021-10-26 修回日期：2022-01-28 本文起止页码：122-130 本文责任编辑：王传清

分析样本,结合计量分析和全文本内容分析方法,对科学数据集在学术论文中的使用特征进行综合考察和分析,并进一步从不同角度分析其对学科发展产生的实际影响。本研究的意义在于,以全学科领域论文作为分析样本,利用规则抽取和自然语言处理技术,从宏观和微观多个层面探讨数据集的使用特征,为科学数据管理和服务提供全新视角的参考,同时也为后续研究提供新的思路。

2 研究方法

2.1 基本思路

本研究全文数据来自于 PubMed Central(PMC)开放获取子集,PMC 是美国国家生物技术信息中心 NCBI(National Center for Biotechnology Information)提供的免费生物医学期刊文献全文数据库[8]。除 PMC 之外,NCBI 还提供 60 余种生物医学数据库以及相关的科研工具,并且通过不同格式为数据库中的科学数据集分配了唯一标识符——登录号(Accession Number)。本文将利用模式匹配的方法,基于自定义规则抽取登录号的方式,在全文中识别并获取数据集的提及信息,将其作为论文对数据集的使用,通过计量分析和内容分析两个维度对数据集使用特征进行分析,并归纳总结科学数据集对于生物医学领域科研和学科发展产生的实际影响力。整体研究思路如图 1 所示:

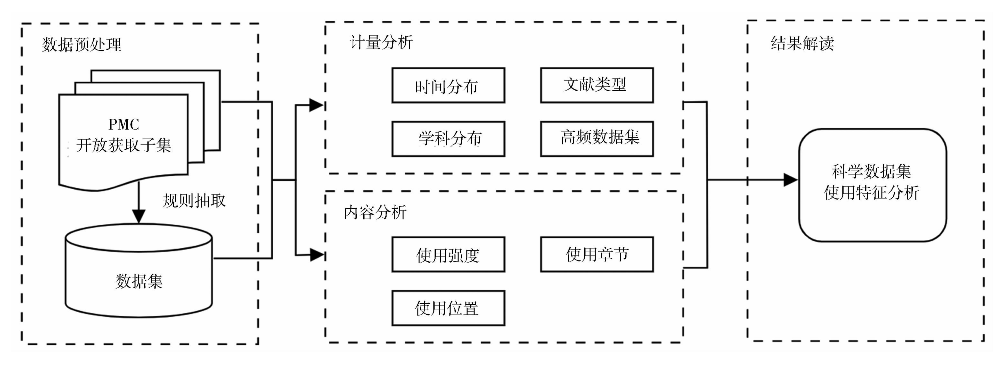


图 1 整体研究思路

2.2 数据获取

本文通过 PMC 提供的 FTP 服务批量下载了 2021 年 5 月 25 日前的文件包,将索引文件合并后获取到文献的基本信息及本地文件位置,PMC 全文数据基于 XML 格式存储,采用美国国立医学图书馆(NLM)的文件类型定义(DTD)标准[9]。最终,共获取到 3 219 908 篇全文文献。

科学数据集的识别采用基于模式匹配的方法,通过正则表达式在全文文本中进行抽取。由于 NCBI 各数据库的登录号规则各异,并且很大一部分数据库的登录号由纯数字组成,无法通过模式匹配方法进行自动抽取。因此,本文选择其中较为常用、格式具备一定规范并且在系统说明中对登录号规则有详细描述的 GEO[10]、RefSeq[11]、SRA[12]、CDD[13] 和 Assembly[14] 5 个数据库中的数据集作为研究对象。其中,GEO 是当前存储规模最大、数据最全面的基因表达数据库,收录了世界各国研究者提交并共享的基因芯片数据和高通量测序数据;RefSeq 是收录基因组、转录本等样本信息的参考序列数据库,为多种生物提供序列的数据信息及相关资料;SRA 数据库主要存储二代测序的原始数据以及与其相关的质控报告;CDD 是蛋白质保守结构域数据库,收集了大量保守结构域序列信息和蛋白质序列信息;组装数据库 Assembly 主要提供已组装基因组结构、相关元数据、组装报告等信息。根据各数据库登录号规则构造的数据集正则表达式如表 1 所示:

表 1 科学数据集登录号示例及正则表达式

<table>
<tr><th>数据库</th><th>正则表达式</th><th>示例</th></tr>
<tr><td>GEO</td><td>G(?:SM|SE|PL|DS)\d+</td><td>GSM22355</td></tr>
<tr><td>RefSeq</td><td>(?:AC|AP|NC|NG|NR|NT|XR|NS)_\d{6}|(?:NM|NP|NW|XM|XP|YP)_\d{6}(?:\d{3})? |(?:ZP_|NZ_[A-Z]{4})\d{8}</td><td>NP_012448</td></tr>
<tr><td>SRA</td><td>SR[PSXR]\d{5,8}</td><td>SRS30216</td></tr>
<tr><td>CDD</td><td>(?:[cC](?:d|l|OG|HL)|[pP](?:HA|LN|TZ|fam)|KOG|MTH|TIGR|smart|LOAD_)\d{4,5}|LOAD_)\d{4,5}</td><td>MTH1747</td></tr>
<tr><td>Assembly</td><td>GC[AF]_\d{9}(?:[.]\d+)?</td><td>GCF_000001735.3</td></tr>
</table>

此外，部分文献还存在如"GSE4357-GSE4380""SRX001799 to SRX001808"等形式的数据集批量使用行为，需要单独构建批量抽取规则，并设置最大抽取阈值为 500，超出则忽略，从而提取出批量使用的数据集登录号。最终经过识别抽取后发现，共有 162 200 篇文献存在本文所涉及 5 个数据库中数据集的使用，数据集总量为 435 920 条，使用次数合计 2 606 552 次，存在数据集使用行为的文献数量占全部文献数量的 5.04%。5 个数据库中被论文使用的数据集数量分布如表 2 所示，其中，RefSeq 数据库中有 238 023 条数据集被使用，约占总量的 55%，说明该数据库中的数据集在生物医学领域得到了较多的关注和使用。

表 2　5 个数据库中被论文使用的数据集数量分布情况

数据库	GEO	RefSeq	SRA	CDD	Assembly
数量/条	86 580	238 023	86 144	13 549	11 624

2.3　计量分析指标

计量分析利用数据集及使用数据集文献的直接指标进行使用特征分析，包括时间分布、文献类型、学科分布和高频数据集。计量分析采用 CountOne 方法[15]，将某一数据集在一篇论文中的多次使用只统计为一次。各项指标的具体分析内容包括：①时间分布：通过对文献数量及使用数据集次数的年度变化趋势进行统计，分析二者随时间的变化规律；②文献类型：使用数据集的文献类型除研究论文和综述以外，还包括报告、简报、评论等类型，对文献类型进行统计分析，发现各类型文献在数据集使用上的特征规律；③学科分布：从刊文期刊所属学科领域角度，探索不同学科领域在数据使用方面的需求差异；④高频数据集：按照使用某一数据集的论文篇数排序，分析高频使用数据集的特征，分析学科研究热点及科研人员使用数据集的习惯和偏好。

2.4　内容分析指标

内容分析利用数据集在文献中提及和使用的详细信息作为间接指标进行使用特征分析，包括使用强度、使用章节和使用位置。内容分析采用 CountX 方法[16]，将某一数据集在一篇论文中出现的使用记录全部纳入分析。各项指标的具体说明如表 3 所示，分析内容包括：①使用强度：采用篇均使用次数作为使用强度，对数据集在论文中的影响力进行评估；②使用章节：将数据使用按照章节类型详细划分 5 个部分，比较分析数据集在论文不同章节的使用情况；③使用位置：较为常见的数据使用位置为正文中的文字描述、表格列出、图片说明等方式，本文采用 8 种数据使用和呈现位置，比较分析数据集在论文中的使用特征。

表 3　数据集使用的内容分析指标分类说明

分析指标	类别或计算方法
使用强度	某个数据集总使用次数/使用该数据集的论文数
使用章节	摘要、引言、数据和方法、结果与讨论、结论
使用位置	正文、表格、图片、参考文献、致谢、附录、脚注、注释

3　结果分析

3.1　计量分析结果

3.1.1　时间分布

1998—2021 年，生物医学领域共有 162 200 篇文献使用了 435 920 条数据集，文献数量和数据集使用量年度分布如图 2 所示。2006 年以后，随着科研范式的转变以及生物信息学、医学信息学等数据驱动型学科的兴起，使用数据集的文献以及数据集的使用数量都开始呈急剧增长的态势。文献数量从 2006 年的 724 篇到 2020 年的 27 279 篇，年均增长率达到 35.5%。数据集的使用次数从 2006 年的 24 783 次到 2020 年的 400 320 次，年均增长率达到 31.5%。科学数据的共享和重用正在深度影响着生物医学相关科研领域的发展，尤其是在近 10 年期间为生物医学开启了崭新的发展阶段。

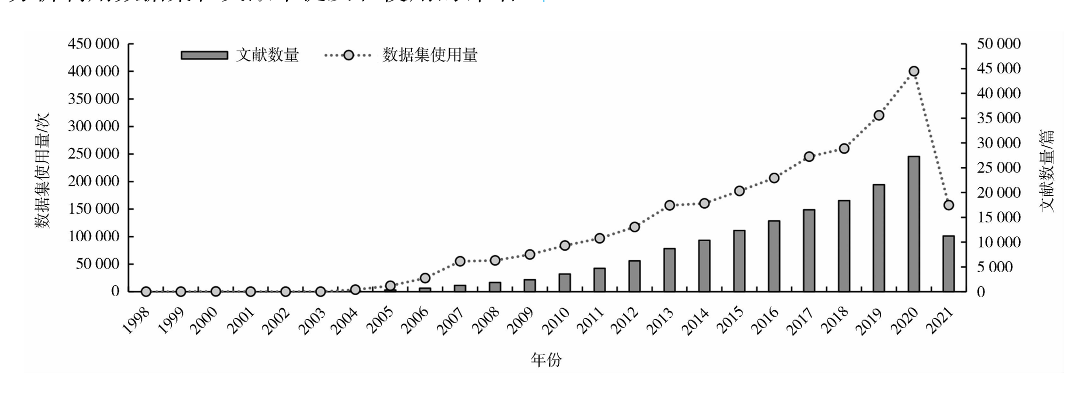


图 2　生物医学领域使用数据集文献及数据集使用次数年度分布情况

3.1.2 文献类型

统计发现有数据集使用行为且标注了类型的文献共 29 种,按照各文献类型数量排序分别为:研究性论文、简报、综述、案例报告、其他、数据论文、通讯、更正、产品综述、摘要、方法论文、社论、系统综述、报告、文章评论、讨论、会议报告、协议、日历、附录、公告、撤稿、章节文章、关注声明、回复、书评、研究快报、描述、新闻。其中,研究性论文约占文献总量的 92%,各种类型文献数量分布如图 3 所示:

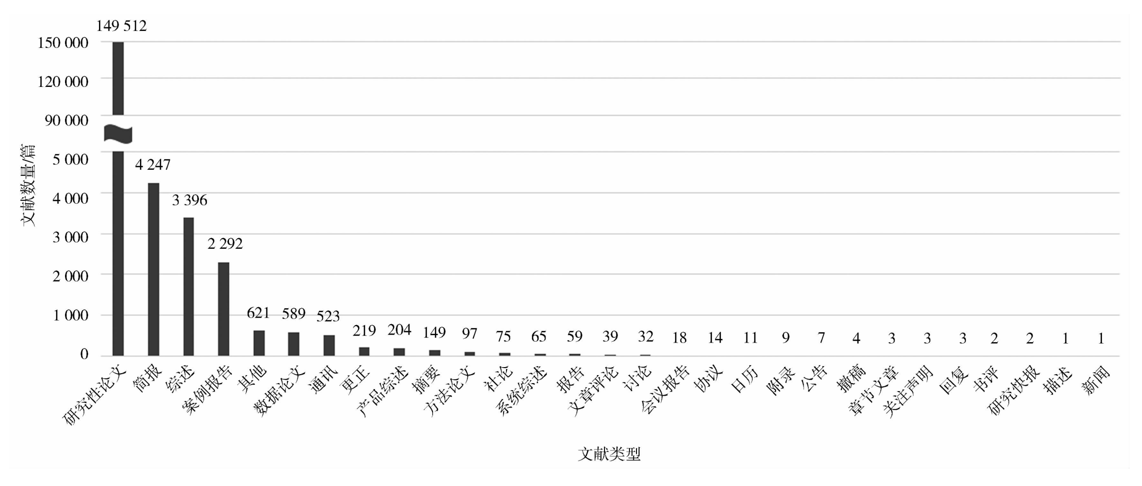


图 3　有数据集使用行为的各种类型文献数量分布情况

除研究性论文外,其余 7 种使用数据集较多的文献类型年度发文量分布如图 4 所示:

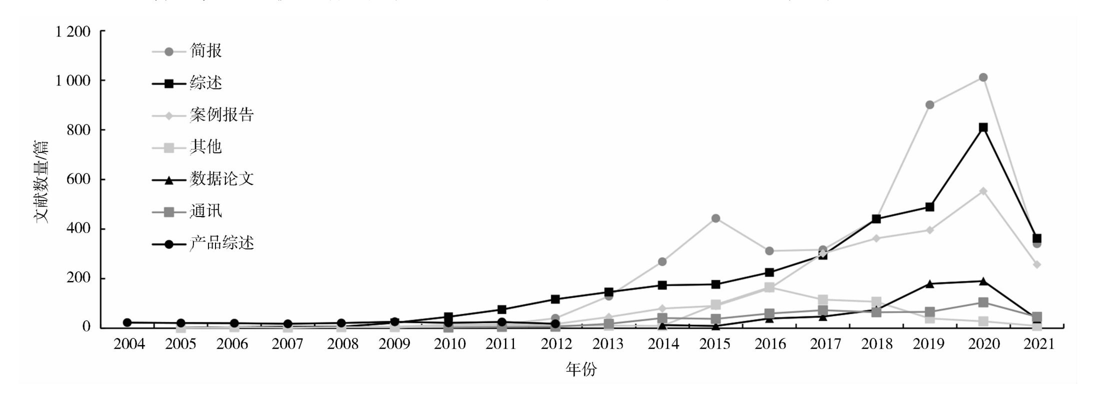


图 4　7 种使用数据集较多的文献类型年度发文量分布情况

由图 4 可知,除研究性论文外,最早使用数据集的文献类型是 2004 年的两篇产品综述,分别使用了 RefSeq 和 CDD 的数据集用于基因数据库构建和蛋白质特异性比对排序软件的开发测试[17－18]。随后,科学数据集开始在简报、综述和案例报告等类型的文献中被使用,尤其是综述文献对数据集的使用逐年平稳增长,说明数据集已经成为一种参与到学科发展历程的科研资料被回顾和使用。此外,2014 年开始出现的数据论文也增长迅速,数据论文作为一种新型学术出版物形式,主要用于描述数据结构、数据处理方法、数据可重用性等内容,数据论文的出现和发展正在积极促进着科学数据的开发和利用[19]。

3.1.3 学科分布

存在数据集使用的文献共发表在 3 127 种期刊上,发文量最多的期刊为《PLOS ONE》,共有 20 931 篇文献存在对数据集的使用。为使研究具备广泛覆盖性并加强分析结果的可解释性,本文排除了发文量较少的期刊,共得到 229 个发文量在 100 篇以上的期刊,总发文量为 131 359 篇,约占文献总量的 81%。本文以中国科学院文献情报中心 2019 年期刊分区表为参考[20],研究并评估前 229 名期刊的研究领域及影响力。经过统计发现,其中共有 181 本 SCI 期刊,Q1 和 Q2 期刊共 120 本,占全部 SCI 期刊的 66%。学科分布及分区见图 5。

从学科分布来看,生物学领域期刊数量占比 56%,生物化学与分子生物学、遗传学、细胞生物学等领域期刊对科学数据集的使用最为频繁。在医学领域,研究与实验、肿瘤学、精神科学等领域期刊较多,是

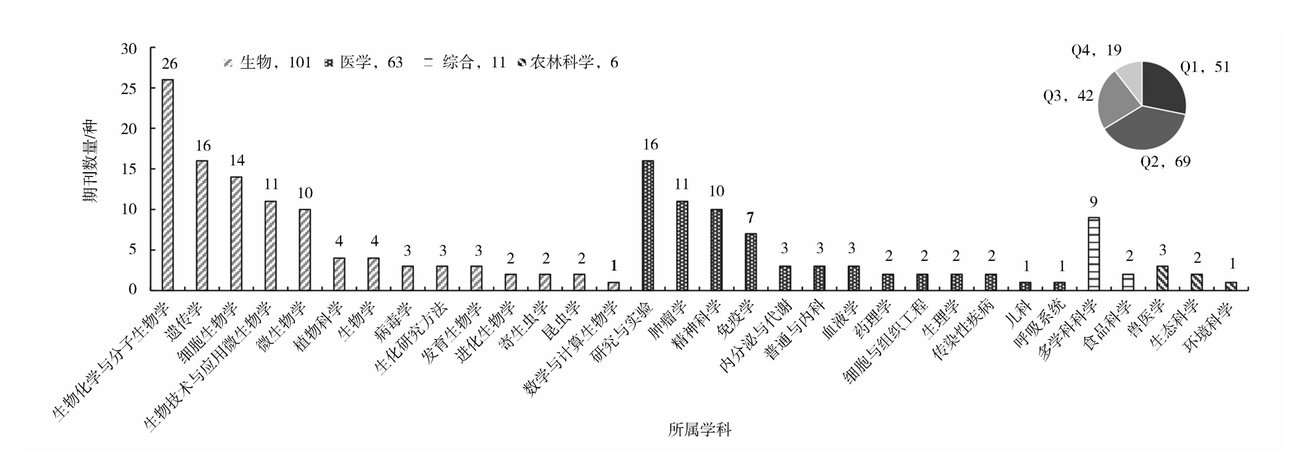
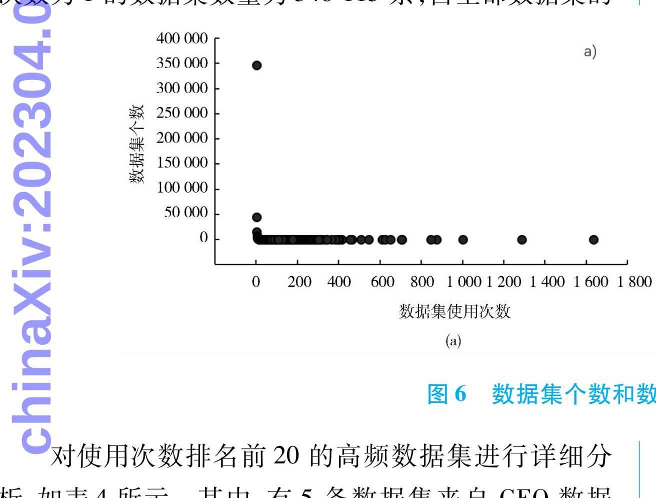


图 5 期刊所属学科分布及分区情况

医学领域使用科学数据集较多的学科。同时，结果中还出现了综合学科、食品科学和农林科学等学科，体现了科学数据集使用的交叉性和跨学科性。

3.1.4 高频数据集

对数据集使用次数进行统计并排序后发现，使用次数为 1 的数据集数量为 346 115 条，占全部数据集的 79%。以数据集使用次数为 X 轴，数据集个数 Y 轴，可以得到图 6 的原始坐标及双对数坐标下的二者关系图。对其进行一元线性回归后得到：log(数据集个数) = 4.59 − 1.91 log(数据集使用次数)，R^2 的值为 0.88，两者呈现出明显的线性关系。结果表明，大量数据集只得到了很少使用，而少数的数据集得到了大量使用。



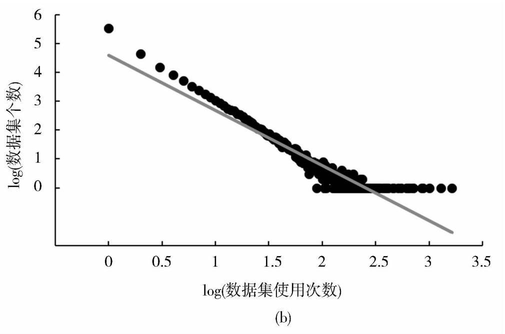


图 6 数据集个数和数据集使用次数关系

对使用次数排名前 20 的高频数据集进行详细分析，如表 4 所示。其中，有 5 条数据集来自 GEO 数据库，其余 15 条数据集都来自 RefSeq 数据库。使用次数最多的“GPL570”数据集是由 Affymetrix 公司提供的商业数据集，该公司是美国著名的生物芯片公司，其余 4 条 GEO 数据集也都出自该公司的芯片产品。从数据集的研究内容和对象来看，围绕肿瘤研究的数据集有 5 条，研究肌动蛋白功能、人类基因组、3 − 磷酸甘油醛脱氢酶的数据集各 3 条，其余数据集与白介素、结核分歧杆菌、大肠杆菌以及新冠病毒的研究相关，从数据集使用频次可以更直观地体现出学科研究的热点和焦点。

表 4 使用次数前 20 名的高频数据集

排序	数据集	物种	次数	排序	数据集	物种	次数
1	GPL570	人类	1 635	11	NC_012920	人类	546
2	NM_002046	人类	1 288	12	NM_013693	小鼠	507
3	NM_001101	人类	1 001	13	NC_000962	结核杆菌	465
4	NM_007393	小鼠	873	14	NM_000546	人类	454
5	NC_000913	大肠杆菌	847	15	NM_008361	小鼠	416
6	GPL96	人类	708	16	NM_031168	小鼠	402
7	NM_008084	小鼠	705	17	GSE31210	人类	395
8	NM_017008	大鼠	652	18	GSE14520	人类	384
9	NC_045512	新冠病毒	627	19	GSE2034	人类	366
10	NM_031144	大鼠	610	20	NM_000600	人类	344

3.2 内容分析结果

3.2.1 使用强度

传统使用频次仅能表明数据集在论文中是否出现，就一篇论文而言，数据集 A 在论文中被反复使用多次，而数据集 B 在论文只被使用一次，则数据集 A 对于该文章的影响力应高于数据集 B，因此本文采用使用强度分析数据集在文献中的使用特征及影响力。从计算结果来看，来自 RefSeq 数据库的数据集“NR_033736”被一篇文献使用了 768 次，成为使用强度最大的数据集[21]。根据数据使用总体情况，本文将科学数据集使用强度划分为 11 个区间，结果如图 7 所示。其中“1”表示数据集在所有使用该数据集的文献中平均被使用 1 次，而“1－2”则表示使用强度大于 1 小于等于 2，以此类推。

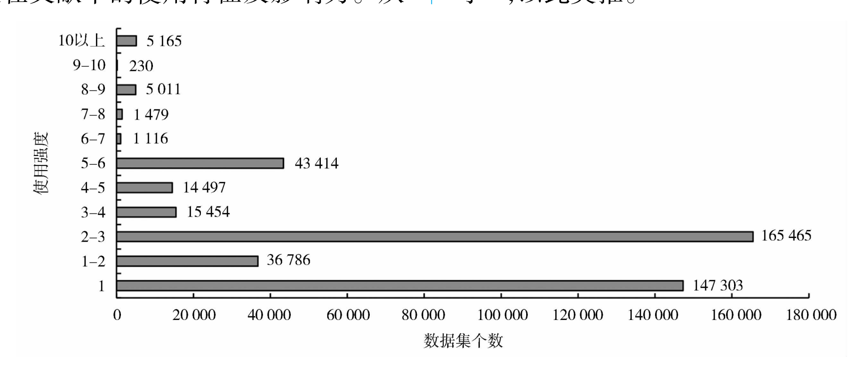


图 7　数据集使用强度分布

由图 7 可以看出，生物医学领域的科学数据集使用强度大多分布在 1－6 之间。其中，使用强度 2－3 之间的最多，其次是 1、5－6、1－2 这几个区间。这与论文引用有着较为明显的区别，相比较而言，科学数据集出现较多高使用强度的现象，表明一条数据集在论文中会被反复使用，贯穿研究的全过程。

3.2.2 使用章节

学术论文各章节的重要性不同，因此在不同章节使用的数据集重要性和影响力也不同。本文结合实证性研究论文 IMRDC 结构将章节划分为 5 个部分[22]，包括摘要、引言、数据和方法、结果与讨论、结论。对于在附录中列出的表格和图片，通过“id”标记可获取其在正文中的使用位置，并将其划分到对应的章节。对于非研究性论文如数据论文、产品综述等文献类型，个别章节无法对应划分到这 5 个章节，则通过人工判读将其划分到功能或位置相近的章节。因为出现这种情况的文献比例较小，对分析结果不会产生太多影响，因此实在无法划分的则排除统计范围之外。使用章节分布结果如图 8 所示：

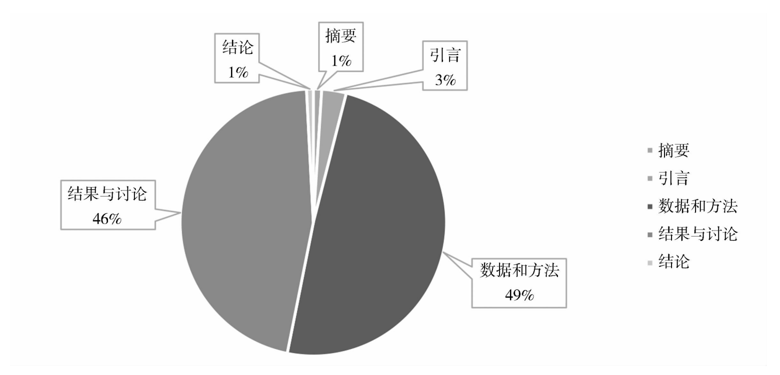


图 8　数据集使用章节分布

由图 8 可知，生物医学领域 49% 的数据集在“数据和方法”部分被使用，其次是“结果与讨论”部分。“摘要”部分是对研究目的、方法、结果和结论的概述，由于篇幅问题，对于所使用的数据不会有过多阐述；“引言”部分会对使用的方法和数据集进行简单的背景介绍，因此会有一定频次的数据集使用；“数据和方法”和“结果与讨论”部分主要围绕数据进行实验分析和结果解读，因此是使用数据集最多的两个部分，约 95% 的数据集使用都出现在这两个章节；“结论”部分会对全文大致流程和结果进行简要总结并对未来工作进行设想，对于具体数据集使用方面的描述较少。总体来看，使用数据集使用的章节分布呈现出极度不平

衡性，这与生物医学领域文献注重实证分析和结果解读有关，并且充分说明科学数据集对于该领域研究的重要性和影响力。

3.2.3 使用位置

与数据集使用章节类似，通过不同位置使用的数据集重要性和影响力也不同。本文将使用位置分为正文、表格、图片、参考文献、致谢、附录、脚注、注释 8 种，正文包括出现在标题、摘要和正文中的数据集。使用位置分布结果如图 9 所示：

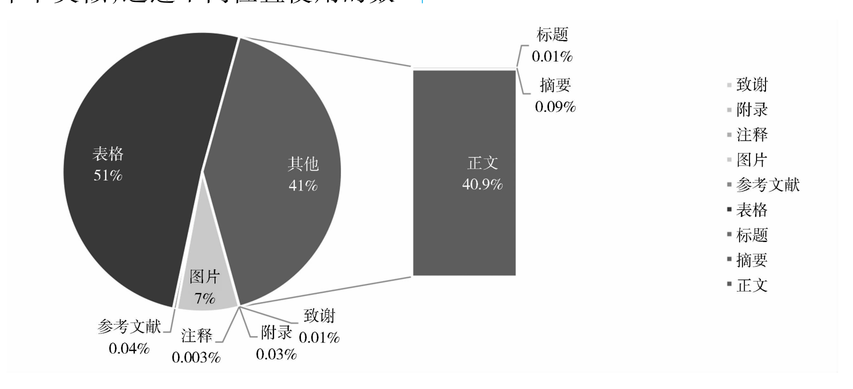


图 9 数据集使用位置分布

由图 9 可知，生物医学领域数据集最常出现的使用位置是通过表格列出，除此之外，正文文字描述、图片说明中也使用数据集较多。对于生物医学领域文献来说，表格和图片的信息同正文具备同样的重要性，在相关研究中应重视表格和图片数据的识别和利用。此外，注释、附录、致谢部分也部分存在数据集的使用。最为值得注意的是参考文献中使用的数据集只占数据集使用总量的 0.04%，这种情况说明在文献中被正式引用的数据集还较少，科学数据集的正式引用问题应得到更多的关注。

4 讨论

由前文的研究结果可以得出如下结论：

（1）科学数据集对生物医学领域科研产生的影响力与日俱增。基于论文数和基于使用强度的统计可以分别代表科学数据集使用的广度和深度。可以设想，一条数据集的影响范围越广，提及该数据集的论文数就会越多。因此，相比较而言数据集的使用广度更能代表其产生的实际影响力，而近 10 余年使用科学数据集论文数量的急剧增长，说明数据集对生物医学领域科研产生的影响力正在与日俱增。同时，数据集的使用强度从另一个角度揭示了科学数据集独特的使用特征。本文研究发现，科学数据集的使用强度要明显高于论文[23]、图书[24]等被引强度，说明科学数据集在论文中较少被当作引言或背景综述提及，而更多的是被实际使用，与论文研究结果紧密相关。

（2）数据出版和高水平期刊促进了科学数据集的开放和共享。从使用科学数据集的文献类型和学科分布可以看出，科学数据集正在逐渐脱离论文，成为一种独立的科研资料，在科学交流过程中发挥着关键作用。目前，数据出版的发展促进了科学数据集的开放和共享，常见的数据出版模式包括数据仓储、数据期刊、数据与论文联合出版 3 种形式，尤其是数据期刊的出现，使得数据论文已经成为近年来发展最为迅速的科学数据发布载体，科学数据正式成为一种可评估、可计量的科研成果产出。从使用数据集论文的学科分布上看，生物医学领域对于数据集的使用非常广泛，数据集产生的影响力正在向综合和交叉学科领域扩展。在进一步对其中的 Q1 区期刊的详细调查中发现，51 个 Q1 区期刊全部都在作者说明或投稿指南中详细说明了数据集的提交要求和提交办法，高水平期刊在开放数据方面的举措无疑加快了数据的共享与重用，推动了科研的发展与进步。

（3）科学数据使用集中在论文的后半部分且正式引用较少。从科学数据集的使用章节和使用方式可以看出，科学数据集出现最多的方式是通过表格列出，其次是正文中提及，在科学数据集使用的相关研究中应注意表格和图片数据的挖掘和利用。而出现最多的章节分别是“数据和方法”“结果与讨论”，这同样与论文、图书等通常被引用在“引言”部分有着明显区别[25-26]。通过结果对比可以发现，不同于其他领域，生物医学论文在“结果与讨论”部分引用论文及使用数据集都较为频繁，说明这一部分是生物医学论文中最为重要的部分，生物医学领域约有 95% 的数据集使用都发生在论文的后半部分。参考文献部分出现的被正式引用的科学数据集比例还较小，说明科学数据集

在论文中仍然以提及等非正式引用方式进行列出或标注,这一方面说明生物医学研究中涉及的数据集数量较多,无法通过参考文献一一列出。另一方面也说明数据正式引用规范还有待发展和完善,数据规范引用对于增强数据价值、提高科研人员共享和重用数据的积极性都具有十分重要的现实意义。

5 总结

本研究以生物医学领域科学数据集为研究对象,通过时间分布、文献类型、学科分布和高频数据集等方面的计量分析,利用数据集及使用数据集文献的直接指标进行使用特征分析,揭示数据集在整个生物医学领域的使用特征规律及产生的影响力;通过使用强度、使用章节、使用位置等方面的全文本内容分析,利用数据集在文献中提及和使用的详细信息作为间接指标进行使用特征分析,从而揭示科学数据集在具体文献中的使用特征及其产生的直接和间接影响力。同前人的研究相比,本研究从宏观和微观两个层面进行分析考量,研究角度更加全面,所得结果也更加完备可靠,可以为科学数据管理和服务工作提供参考依据。首先,要进一步推进科学数据引用标准规范的建立,提高科学数据库对于唯一标识符、版本号的分配和管理,规范的数据引用对于提高科研工作者的数据引用意识、追溯数据使用情况都具有非常重要的意义;其次,科学数据库建设要具备专业性、及时性和开放性,专业性的数据库具备更强的吸引力,数据要由专业运维团队及同行评议专家进行及时更新维护,通过多渠道资金优化配置保证数据的免费和开放访问是科学数据库长远建设发展的保证;最后,高校和图书馆要加强科学数据人才培养,包括数据管理研究型人才、数据分析型人才、数据监管型人才等,满足飞速发展的科学数据管理和服务需求。

当然,本文研究也存在着一些不足:一方面,由于科学数据集识别和抽取方法的局限,本文只针对 NCBI 旗下登录号较为规范的 5 个数据库中的科学数据集进行抽取和研究,研究范围存在一定局限性;另一方面,本文只从论文角度进行研究和分析,并未深入到数据集的元数据和内容信息,并且将数据集的提及等同于使用,而并未对数据集的使用意图进行进一步研究和划分,揭示层次还较浅,分析深度还有待进一步加强。在今后工作中,将继续提高科学数据集识别的范围和准确性,从更细粒度的角度继续分析挖掘科学数据集的使用特征和影响力。

参考文献:

[1] 屈宝强，王凯. 科学数据引用现状和研究进展[J]. 情报理论与实践,2016,39(5):118－138.
[2] 朱少强，邱均平. 文献计量与内容分析——文献群中隐含信息的挖掘[J]. 图书情报工作,2005(6):19－23.
[3] BELTER C W, BROWMAN H I. Measuring the value of research data: a citation analysis of oceanographic data sets[J]. Plos one, 2014,9(3):e92590.
[4] 焦红，杨波，周琪. 生物医学领域科学数据集复用特征研究[J]. 情报理论与实践,2021,44(9):90－96.
[5] 王曰芬，路菲，吴小雷. 文献计量和内容分析的比较与综合研究[J]. 图书情报工作,2005,49(9):72－75.
[6] 王雪，马胜利，佘曾溧，等. 科学数据的引用行为及其影响力研究[J]. 情报学报,2016,35(11):1132－1139.
[7] 李龙飞，余厚强，尹梓涵，等. 替代计量学视角下科学数据集价值的定量测度研究[J]. 情报理论与实践,2020,43(9):47－52,71.
[8] 沈锡宾，吕小东，郝秀原，等. PubMed Central 简介及其对期刊的评估和收录[J]. 中国科技期刊研究,2006,17(5):866－868.
[9] 沈锡宾，顾佳，包婧玲，等. 美国 NLM DTD 3.0 期刊存储和交换标签集中参考文献的标记解读[J]. 中国科技期刊研究,2013,24(2):233－237.
[10] NCBI. Gene expression omnibus [EB/OL]. [2021－07－12]. https://www.ncbi.nlm.nih.gov/geo/.
[11] NCBI. Reference sequence database [EB/OL]. [2021－07－12]. https://www.ncbi.nlm.nih.gov/refseq/.
[12] NCBI. Sequence read archive [EB/OL]. [2021－07－12]. https://trace.ncbi.nlm.nih.gov/Traces/sra/.
[13] NCBI. Conserved domains database [EB/OL]. [2021－07－12]. https://www.ncbi.nlm.nih.gov/cdd/.
[14] NCBI. Assembly [EB/OL]. [2021－07－12]. https://www.ncbi.nlm.nih.gov/assembly/.
[15] WAN X, LIU F. WL－index: leveraging citation mention number to quantify an individual's scientific impact[J]. Journal of the American Society for Information Science & Technology, 2014, 65(12):2509－2517.
[16] DING Y, LIU X, GUO C, et al. The distribution of references across texts: some implications for citation analysis[J]. Journal of informetrics,2013,7(3):583－592.
[17] WANG B B, BREN DE L V. The asrg database: identification and survey of arabidopsis thaliana genes involved in pre-mRNA splicing [J]. Genome biology,2004,5(12):1－23.
[18] MEEREIS F, KAUFMANN M. Pcogr: phylogenetic cog ranking as an online tool to judge the specificity of cogs with respect to freely definable groups of organisms[J]. BMC bioinformatics, 2004, 5(1):150－150.
[19] 屈宝强，王凯. 数据论文的出现与发展[J]. 图书与情报,2015(5):1－8.

[20] 中国科学院文献情报中心. 中国科学院文献情报中心期刊分区表[EB/OL]. [2021 - 07 - 12]. http://www.fenqubiao.com/.
[21] LI J, JIN K, LI M, et al. A host cell long noncoding RNA nr_033736 regulates type I interferon - mediated gene transcription and modulates intestinal epithelial anti - cryptosporidium defense[J]. Plos Pathogens,2021,17(1):e1009241.
[22] LIN L, EVANS S. Structural patterns in empirical research articles: a cross - disciplinary study[J]. English for specific purposes,2012,31(3):150 - 160.
[23] 胡志刚. 全文引文分析方法与应用[M]. 北京:科学出版社,2017.
[24] 章成志, 李卓, 赵梦圆, 等. 基于引文内容的中文图书被引行为研究[J]. 中国图书馆学报,2019,45(3):96 - 109.
[25] 张梦莹, 卢超, 郑茹佳, 等. 用于引文内容分析的标准化数据集构建[J]. 图书馆论坛,2016(8):48 - 53.
[26] CHI P S. Differing disciplinary citation concentration patterns of book and journal literature? [J]. Journal of informetrics,2016,10(3):814 - 829.

作者贡献说明:

杨宁:资料搜集,实验验证及初稿撰写;
张志强:指导论文修改,凝练论文研究点。

Research on the Use Characteristics of Scientific Datasets Combined with Quantitative Analysis and Content Analysis

Yang Ning[1,2] Zhang Zhiqiang[1,2]
[1] Chengdu Library and Information Center, Chinese Academy of Sciences, Chengdu 610041
[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

[Purpose/significance] This paper analyzes the use characteristics of scientific datasets from the perspective of quantitative analysis and content analysis, quantitatively evaluates the impact of scientific datasets on discipine development, and provides references for scientific data management services and policy research. **[Method/process]** Methods of text mining and bibliometric were used to analyze the full - text literature in PubMed Central, this study comprehensively investigated the use of scientific datasets from 7 aspects such as time distribution and use intensity, and on this basis, evaluated the actual impact of scientific datasets on discipline development. **[Result/conclusion]** The research results show that the influence of scientific datasets on scientific research in the biomedical field is increasing with each passing day. Data publishing and high - level journals promote the opening and sharing of scientific datasets. The use of scientific datasets is concentrated in the second half of the paper and there are few formal references. The corresponding standards and specifications need to be further strengthened.

Keywords: quantitative analysis content analysis scientific dataset use characteristics